Meta-Analysis for Genome-wide Association Studies

Yves Chretien
Department of Statistics, Harvard University

Outline

- Introduction
- Current approaches
- A meta-analysis framework
- Continuing work

Introduction

- Genome-wide association studies (GWAS's)
  - Useful for common, "complex" disease
  - Multiple studies being performed for certain diseases
  - Typical study:
    - ~1000 cases, ~1000 controls
    - ~100K-500K markers (SNPs) spanning genome

Introduction

- How to best pool information across studies?
  - For some diseases, multiple studies now available
  - Meta-analysis framework needed
  - For linkage studies, GSMA (Genome Scan Meta-Analysis – Wise et al 1999, Levinson et al 2003)

Introduction

- Thus far, meta-analysis methods for association studies somewhat *ad hoc*

- Systematic strategies have been proposed (Evangelou et al 2007). However, currently limited by suboptimal overlap of marker sets between studies

Criteria for ideal meta-analysis method

- Should be a generalization of the analysis used for a single GWAS
- As in GSMA, should allow inference on SNPs regardless of whether genotyped in each study
- Should rely on known LD structure in a region for inference about SNPs not genotyped in a given study
- Should be the optimal analysis in the ideal case where all SNPs typed in all studies

Single GWAS

	Case haplotypes	Control haplotypes
A		
a		

$D = \{\text{case haplotype}\}$; $\bar{D} = \{\text{control haplotype}\}$; alleles = A/a

$$P(A \mid D) = \frac{e^{\delta+\theta}}{1+e^{\delta+\theta}},\ P(A \mid \bar{D}) = \frac{e^{\delta}}{1+e^{\delta}}$$

$$\mathrm{H}_0\text{: } \theta = 0$$

Score test = standard 1-df chi-square test for 2x2 table

Two GWAS's

Study #1

	Case haplotypes	Control haplotypes
A		
a		

Study #2

	Case haplotypes	Control haplotypes
A		
a		

$$P\left(A \mid D\right) = \frac{e^{\delta_1 + \theta}}{1 + e^{\delta_1 + \theta}}$$

$$P\left(A \mid \bar{D}\right) = \frac{e^{\delta_1}}{1 + e^{\delta_1}}$$

$$P\left(A \mid D\right) = \frac{e^{\delta_2 + \theta}}{1 + e^{\delta_2 + \theta}}$$

$$P\left(A \mid \bar{D}\right) = \frac{e^{\delta_2}}{1 + e^{\delta_2}}$$

$$\mathrm{H}_0\text{: } \theta = 0$$

Score test = Mantel-Haenszel test

Two GWAS's

Study #k (k = 1,2)

	Case haplotypes	Control haplotypes	
A	n_{A1k}	n_{A2k}	$n_{A.k}$
a	n_{a1k}	n_{a2k}	$n_{a.k}$
	$n_{.1k}$	$n_{.2k}$	$n_{..k}$

$$p_{1k} = \frac{n_{A1k}}{n_{.1k}}$$

$$p_{2k} = \frac{n_{A2k}}{n_{.2k}}$$

$$\overline{p}_k = \frac{n_{A.k}}{n_{..k}}$$

$$\psi_{MH} = \frac{\left(\left| \sum_{k=1}^{2} \frac{n_{.1k} n_{.2k}}{n_{..k}} (p_{1k} - p_{2k}) \right| - \frac{1}{2} \right)^2}{\sum_{k=1}^{2} \frac{n_{.1k} n_{.2k}}{(n_{..k} - 1)} \overline{p}_k \overline{q}_k}$$

Two GWAS's

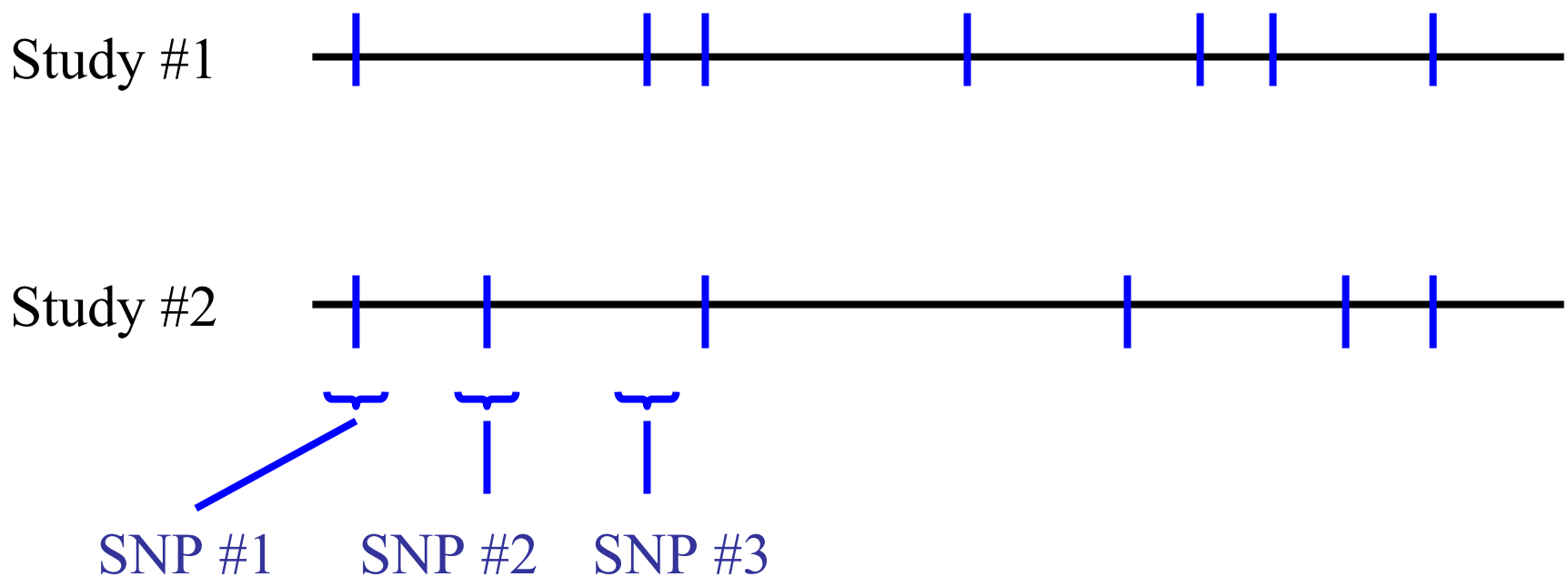
Any two GWAS's will use overlapping but non-identical SNP maps; thus, not every SNP genotyped in both studies

Still want a measure of statistical significance for each individual SNP, taking into account all evidence ("direct" and "indirect") from both studies

SNP maps for two GWAS's

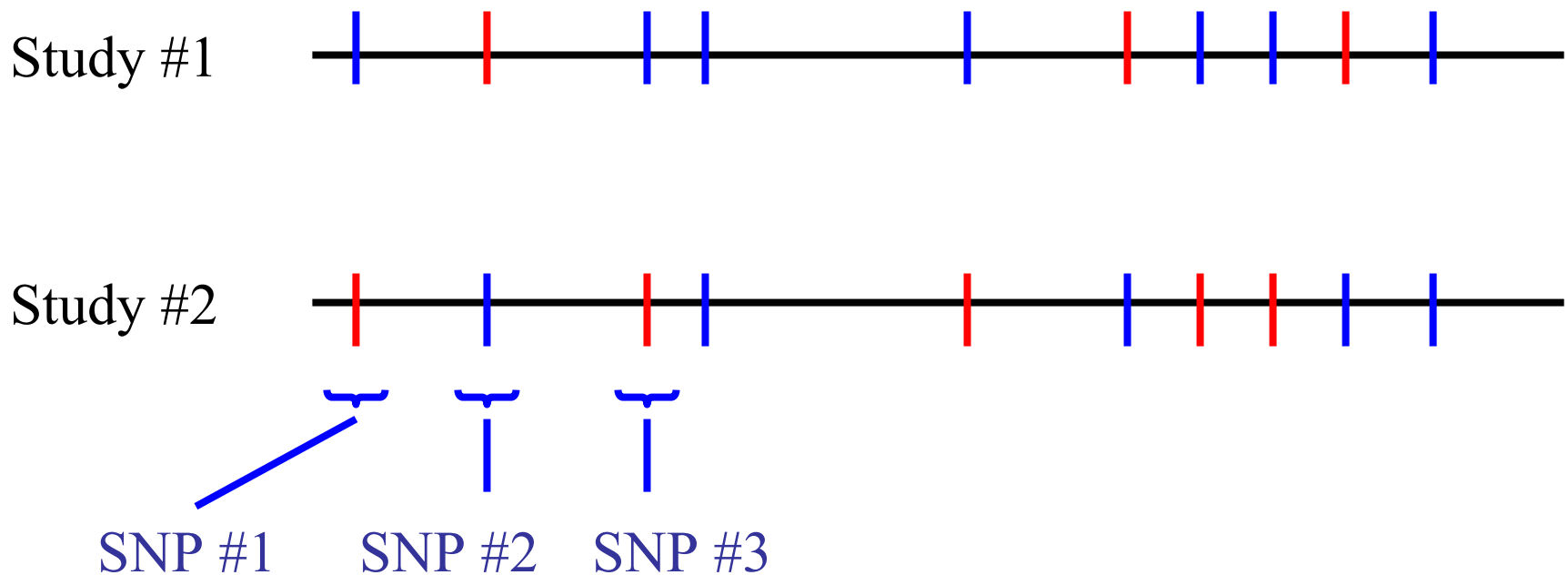
Some SNPs genotyped in both studies,

some SNPs genotyped in one study only



SNP maps for two GWAS's

<u>Strategy:</u> impute allele frequencies for all "missing" SNPs, then analyze as if imputed data were observed data



Imputing missing data

- HapMap provides independent sample of haplotypes that can be used to impute missing alleles in a study.
- The HapMap has been used for testing at "hidden" sites in single-study analyses…
  - Nicolae (2006b) introduces TUNA (Testing Untyped Alleles)
  - Zaitlen et al (2007) introduce WHAP (Weighted sum of Haplotype frequency difference)
- We extend this concept to meta-analysis problems
- **Key assumption:** same LD in HapMap population as in study population!

How to impute?

If a given SNP is not genotyped in some study, we need to impute the allele frequencies in the cases and controls of that study.

<u>Step 1:</u> choose best multimarker predictor in HapMap, e.g. using Tagger (de Bakker et al, 2005)

<u>Step 2:</u> impute allele frequencies in cases & controls

<u>Step 3:</u> plug estimated frequencies into Ψ_{MH} and treat as observed frequencies

How to impute?

Example: SNP #1 not genotyped in study; its frequency in cases imputed using HapMap data

haplotypes			Frequency in HapMap
SNP #1	SNP #2	SNP #3	
T	T	C	0.30
T	T	A	0.30
C	T	A	0.10
C	G	C	0.20
C	G	A	0.10

$$\hat{p}_{T\bullet\bullet} = p_{\bullet TC} + \frac{0.30}{0.30 + 0.10} p_{\bullet TA}$$

Type I error rate

HapMap samples are used to impute allele frequencies for **both** cases and controls

Imputation thus performed as though H_0 were true

Therefore, the null distribution of Ψ_{MH} is the **same**, regardless of whether real or imputed data are used

Power

Clearly power is lost when we impute rather than observe all genotypes.

Rely on assumption that underlying LD in HapMap is identical to that in study subjects

Under this assumption, relative efficiency of procedure (to complete-data case) depends on ability of proxy markers to predict hidden site (measured by multi-marker R^2)

Hypothesis testing using Ψ_{MH}

H_0: $\theta = 0$

H_1: $\theta \neq 0$ (common across studies)

To derive the alternative distribution of Ψ_{MH} we condition on the numbers of cases and controls in each study…

$$\text{Let } \Delta_k = P\left(A \mid D\right) - P\left(A \mid \bar{D}\right) \text{ in study } k$$

Null distribution of Ψ_{MH} : All genotypes observed

$$(p_{1k} - p_{2k}) \sim N\left(0, \frac{n_{..k}}{n_{.1k} n_{.2k}} \bar{p}_k \bar{q}_k\right)$$

$$\Downarrow$$

$$\frac{\sum_{k=1}^{2} \frac{n_{.1k} n_{.2k}}{n_{..k}} (p_{1k} - p_{2k})}{\sum_{k=1}^{2} \frac{n_{.1k} n_{.2k}}{n_{..k}} \bar{p}_k \bar{q}_k} \sim N(0,1)$$

Alternative distribution of Ψ_{MH} : All genotypes observed

$$\boxed{(p_{1k} - p_{2k}) \approx N\left(\Delta_k, \frac{n_{..k}}{n_{.1k} n_{.2k}} \overline{p}_k \overline{q}_k\right)}$$

$$\Downarrow$$

$$\boxed{\frac{\sum_{k=1}^{2} \frac{n_{.1k} n_{.2k}}{n_{..k}} (p_{1k} - p_{2k})}{\sum_{k=1}^{2} \frac{n_{.1k} n_{.2k}}{n_{..k}} \overline{p}_k \overline{q}_k} \approx N\left(\frac{\sum_{k=1}^{2} \frac{n_{.1k} n_{.2k}}{n_{..k}} \Delta_k}{\sum_{k=1}^{2} \frac{n_{.1k} n_{.2k}}{n_{..k}} \overline{p}_k \overline{q}_k}, 1\right)}$$

Alternative distribution of Ψ_{MH} : Some genotypes imputed

Suppose marker A not genotyped in (at least) one of the studies; instead we choose some "proxy markers" nearby. We estimate frequency of A allele among cases in that study as

$$\hat{p}_1 = \sum_h \left(\frac{f_{hA}}{f_{hA} + f_{ha}} \right) p_{h1}$$

where

f_{hA} = HapMap frequency of hA haplotype

p_{h1} = frequency of haplotype h among cases

and the summation is over all observed haplotypes generated by the proxy markers

Alternative distribution of Ψ_{MH} : Some genotypes imputed

To find the mean and variance of this estimator, we need to know the distribution of haplotype frequencies given the frequency of allele A, since

$$\sum_h \left(\frac{f_{hA}}{f_h} \right) p_{h1} = \sum_h \left(\frac{f_{hA}}{f_h} \right) \left\{ P\left(h \mid A, D\right) p_1 + P\left(h \mid a, D\right)\left(1 - p_1\right) \right\}$$

These could also be estimated using the HapMap, under the assumption that it accurately represents LD in study population

$$\hat{P}\left(h \mid A, D\right) = \left(\frac{f_{hA}}{f_A} \right) \quad , \quad \hat{P}\left(h \mid a, D\right) = \left(\frac{f_{ha}}{f_a} \right)$$

Alternative distribution of Ψ_{MH} : Some genotypes imputed

This allows us to estimate, for such a study,

$$E\left(\hat{p}_1 - \hat{p}_2\right) \approx \left(p_1 - p_2\right) \sum_h \left(\frac{f_{hA}}{f_h}\right)\left(\frac{f_{hA}}{f_A} - \frac{f_{ha}}{f_a}\right) \triangleq \left(p_1 - p_2\right) \lambda_1$$

$$Var\left(\hat{p}_1 - \hat{p}_2\right) \approx \overline{p}\overline{q}\left(\frac{n_{..}}{n_{.1}n_{.2}}\right) \sum_h \left(\frac{f_{hA}}{f_h}\right)^2 \left(\frac{f_{hA}}{f_A} - \frac{f_{ha}}{f_a}\right)^2$$

$$\triangleq \overline{p}\overline{q}\left(\frac{n_{..}}{n_{.1}n_{.2}}\right) \lambda_2$$

Alternative distribution of Ψ_{MH} : Some genotypes imputed

Now define the following for each study *k:*

$$\lambda_{1k} = \begin{cases} 1, & \text{if maker } A \text{ genotyped in study } k \\ \sum_h \left(\frac{f_{hA}}{f_h} \right) \left(\frac{f_{hA}}{f_A} - \frac{f_{ha}}{f_a} \right), & \text{if not} \end{cases}$$

$$\lambda_{2k} = \begin{cases} 1, & \text{if maker } A \text{ genotyped in study } k \\ \sum_h \left(\frac{f_{hA}}{f_h} \right)^2 \left(\frac{f_{hA}}{f_A} - \frac{f_{ha}}{f_a} \right)^2, & \text{if not} \end{cases}$$

$$\left(\tilde{p}_{1k}, \tilde{p}_{2k} \right) = \begin{cases} \left(p_{1k}, p_{2k} \right), & \text{if maker } A \text{ genotyped in study } k \\ \left(\hat{p}_{1k}, \hat{p}_{2k} \right), & \text{if not} \end{cases}$$

Alternative distribution of Ψ_{MH} : Some genotypes imputed

This allows us to estimate

$$\boxed{\frac{\sum_{k=1}^{2} \frac{n_{.1k} n_{.2k}}{n_{..k}} \frac{(\tilde{p}_{1k} - \tilde{p}_{2k})}{\lambda_{2k}}}{\sum_{k=1}^{2} \frac{n_{.1k} n_{.2k}}{n_{..k}} \frac{\overline{p}_k \overline{q}_k}{\lambda_{2k}}} \approx N\left(\frac{\sum_{k=1}^{2} \frac{n_{.1k} n_{.2k}}{n_{..k}} \left(\frac{\lambda_{1k}}{\lambda_{2k}} \right) \Delta_k}{\sum_{k=1}^{2} \frac{n_{.1k} n_{.2k}}{n_{..k}} \frac{\overline{p}_k \overline{q}_k}{\lambda_{2k}}}, 1 \right)}$$

Power of test using imputed data

Therefore, we can define a new test statistic, $\widetilde{\Psi}_{MH}$, that is based on the observed plus the imputed data.

Under H_1, $\widetilde{\Psi}_{MH} \sim$ non-central χ^2 with a non-centrality parameter that can be approximated as

$$\left(\frac{\sum_{k=1}^{2} \frac{n_{.1k} n_{.2k}}{n_{..k}} \left(\frac{\lambda_{1k}}{\lambda_{2k}} \right) \Delta_k}{\sum_{k=1}^{2} \frac{n_{.1k} n_{.2k}}{n_{..k}} \frac{\overline{p}_k \overline{q}_k}{\lambda_{2k}}} \right)^2$$

Continuing work: empirical validation of method

All of these heuristics are based on assumption that LD in the HapMap accurately reflects LD in the study populations

Next step: Monte Carlo simulation to test validity of this heuristic regarding power. Simulate case & control genotypes at a set of markers; compare power using all genotypes with power when one study's genotypes are imputed

Continuing work: Bayesian modeling

- p-values not correct if markers for one study had been chosen based on results from another study in the meta-analysis (e.g., Lohmueller et al 2003); posterior probability distributions not problematic in this situation

- Bayesian model could allow differences in LD between HapMap and study populations

Acknowledgments

Jun Liu (Harvard)

David Altshuler & Mark Daly Groups (Broad Institute)

NSF Conference Organizers

References

de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D (2005) Efficiency and power in genetic association studies. *Nature Genetics*. **37**: 1217-1223

Evangelou E, Maranganore DM, Ioannidis JP (2007). Meta-analysis in genome-wide association datasets: strategies and application in Parkinson disease. *PLoS ONE,* 2:e196

Levinson, DF, Levinson MD, Segurado R, Lewis CM (2003) Genome scan meta-analysis of schizophrenia and bipolar disorder, part I: methods and power analysis. *Am. J. Hum. Genet,* **73**, 17–33

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* **33**: 177–82.

Nicolae DL (2006a) Quantifying the amount of missing information in genetic association studies. *Genetic Epidemiology,* **30**(8): 703-717

Nicolae DL (2006b) Testing Untyped Alleles (TUNA) – applications to genome-wide association studies. *Genetic Epidemiology,* **30**(8): 718-27

Wise, LH, Lanchbury JS, Lewis CM (1999) Meta-analysis of genome searches. *Ann. Hum. Genet*, . **63**, 263–272

Zaitlen N, Kang HM, Eskin E, Halperin E (2007) Leveraging the HapMap Correlation Structure in Association Studies. *Am. J. Hum. Genet*, **80**(4): 683-691